Unsupervised Learning by Examples: On-line versus Off-line

C. Van den Broeck and P. Reimann*
Limburgs Universitair Centrum, B-3590 Diepenbeek, Belgium
(Received 31 October 1995)

We study both on-line and off-line unsupervised learning from p random patterns which are uniformly distributed on the N-sphere with the exception of a single symmetry breaking orientation $\boldsymbol{B}$, along which they may be arbitrarily distributed. Supervised learning from the same kind of patterns is included as a special case. In the thermodynamic limit $N \to \infty$ with $\alpha = p/N$ fixed we calculate the overlap $R(\alpha) = \boldsymbol{B} \cdot \boldsymbol{J}/|\boldsymbol{J}||\boldsymbol{B}|$ between the unknown "true" $\boldsymbol{B}$ and the optimal "Bayes" hypothesis $\boldsymbol{J}$ with particular emphasis on the small and large α asymptotics and the phenomenon of retarded learning. Finally, we identify a cost function whose minimum reproduces the off-line Bayes overlap.

PACS numbers: 87.10.+e, 02.50.–r, 05.20.–y

Over the last decade, networks that can learn by examples have been investigated theoretically by applying a formalism similar to that of statistical mechanics. Typically, the network is trained off-line by minimization of a cost function which incorporates the information about the whole training set [1–8]. More recently, very interesting results have been obtained for on-line learning, a step-wise learning procedure in which the training examples are presented only once and in a sequential order [9–12]. It turns out that the analytic calculations for this case are rather simple, with an optimal performance often close to that for off-line learning, while its practical implementation is straightforward. Most off-line and all on-line theoretical calculations have been restricted to the case of supervised learning, a scenario in which a set of training patterns together with their corresponding classifications are given. In this Letter, we present results for *unsupervised* learning, both on-line and off-line, including the supervised scenario as a special case. We mainly focus on Bayes learning, defined as the situation in which the available information is used optimally, because the on-line and off-line versions are then closely related and give similar results. Additionally, we identify a cost function with a unique minimum that reproduces the result of off-line Bayes learning. It sheds light on the relation between the on-line and off-line Bayes solutions.

We consider the following unsupervised problem. Patterns $\boldsymbol{\xi}$ are sampled independently from a nonuniform distribution $P_{\boldsymbol{B}}(\boldsymbol{\xi})$ with a single symmetry breaking orientation $\boldsymbol{B}$. Without loss of generality we can choose the length of $\boldsymbol{\xi}$, $\boldsymbol{B}$, and other N-dimensional vectors equal to $\sqrt{N}$, with N the dimensionality of the pattern space. Furthermore, the axial symmetry around $\boldsymbol{B}$ implies that $P_{\boldsymbol{B}}(\boldsymbol{\xi})$ is proportional to $e^{-U(b)}\delta(\boldsymbol{\xi}^2 - N)$, where $U(b)$ describes the modulation of the pattern density as a function of the overlap $b = \boldsymbol{\xi} \cdot \boldsymbol{B}/\sqrt{N}$. In the limit $N \to \infty$, the distribution $\mathcal{P}(b)$ of this overlap takes on the following form:

$$\mathcal{P}(b) = \frac{\mathcal{N}}{\sqrt{2\pi}}\, e^{-b^2/2 - U(b)}, \tag{1}$$

where $\mathcal{N}$ is a normalization constant. Our aim is to construct an estimate $\boldsymbol{J}$ of the unknown orientation $\boldsymbol{B}$ on the basis of a training set of p patterns $\{\boldsymbol{\xi}\}_{\mu=1}^{p}$ sampled independently from $P_B(\boldsymbol{\xi})$ under the assumption that the function $U(b)$ in (1) is exactly known. The quality of a hypothesis $\boldsymbol{J}$ will be measured by its overlap $R = \boldsymbol{J} \cdot \boldsymbol{B}/N$ with the true orientation $\boldsymbol{B}$. The case of supervised learning from a teacher, coinciding with $\boldsymbol{B}$, and providing, in addition, a binary classification $\xi_0^\mu = \mathrm{sgn}[f(\boldsymbol{\xi}^\mu \cdot \boldsymbol{B}/\sqrt{N})]$ of each pattern $\boldsymbol{\xi}^\mu$, can be transformed into the unsupervised scenario by estimating the orientation of $\boldsymbol{B}$ from the "aligned" patterns $\boldsymbol{\xi}^\mu \xi_0^\mu$. This transformation is exact provided $f(\lambda)$ is an odd function, i.e., the additional knowledge of the classifications ξ_0^μ does not allow a better guess of $\boldsymbol{B}$ than the $\boldsymbol{\xi}^\mu \boldsymbol{\xi}_0^\mu$ alone [13]. The distribution of the aligned patterns is given by

$$\tilde{\mathcal{P}}(b) = [\mathcal{P}(b) + \mathcal{P}(-b)]\Theta(f(b)), \tag{2}$$

where $\Theta(x)$ is the Heaviside function.

In on-line learning, the estimated orientation $\boldsymbol{J}$ is updated, upon presentation of a new pattern $\boldsymbol{\xi}$, to a new one $\boldsymbol{J}'$ as follows:

$$\boldsymbol{J}' = \sqrt{N/(\boldsymbol{J} + F\boldsymbol{\xi}/\sqrt{N})^2}\,(\boldsymbol{J} + F\boldsymbol{\xi}/\sqrt{N}). \tag{3}$$

The prefactor on the right hand side (r.h.s.) of Eq. (3) guarantees that $|\boldsymbol{J}'|^2 = N$. The meaning of the amplitude F, with which the new pattern $\boldsymbol{\xi}$ contributes to the reorientation of $\boldsymbol{J}$, can be clarified by multiplying both sides in Eq. (3) by $\boldsymbol{\xi}/\sqrt{N}$. Keeping terms to dominant order in N, one concludes that $F = \lambda - t$, where λ and t are the overlap of the pattern $\boldsymbol{\xi}$ after and prior to "learning": $\lambda = \boldsymbol{J}' \cdot \boldsymbol{\xi}/\sqrt{N}$ and $t = \boldsymbol{J} \cdot \boldsymbol{\xi}/\sqrt{N}$.

To identify the optimal choice of F, corresponding to "Bayes on-line learning," we multiply both sides in Eq. (3) with $\boldsymbol{B}$, and find that the increase in overlap $\Delta R = (\boldsymbol{J}' - \boldsymbol{J}) \cdot \boldsymbol{B}/N$ is maximal for $F = b/R - t$ with $b = \boldsymbol{B} \cdot \boldsymbol{\xi}/\sqrt{N}$. The latter quantity, however, is not available. To realize, on average, a maximal increase of the overlap, the best one can do is to use the conditional

 0031-9007/96/76(12)/2188(4)$10.00

average of b, calculated on the basis of all the available information, in this case the value of the overlap t prior to learning,

$$F_{\rm opt}(R,t) = R^{-1}\langle b \mid t\rangle - t\,, \tag{4}$$

with $\langle b \mid t\rangle = \int bP(b \mid t)\,db$. With this choice of the learning amplitude, one finds in the continuous limit $p \to \infty$, $N \to \infty$ with $\alpha = p/N$ finite, that the Bayes overlap $R = R(\alpha)$ becomes a self-averaging quantity and its evolution $dR/d\alpha$ is given by $(R/2)\int F_{\rm opt}(R,t)^2 P(t)\,dt$. In order to complete these formulas, we invoke the cylindrical symmetry around the $\boldsymbol{B}$ axis. It follows that $P(t \mid b)$ is a Gaussian distribution with average $\langle t \mid b\rangle = bR$ and dispersion $1 - R^2$. One thus obtains the joint probability distribution $P(b,t) = P(t \mid b)\mathcal{P}(b)$ and from it the distributions $P(t) = \int P(b,t)\,db$ and $P(b \mid t) = P(b,t)/P(t)$. The evolution equation for R can then be rewritten explicitly as follows:

$$\frac{dR^2}{d\alpha} = (1 - R^2)\int \mathcal{D}t\,\frac{\{\int \mathcal{D}t'\,t'\mathcal{N}\,e^{-U(Rt+\sqrt{1-R^2}\,t')}\}^2}{\int \mathcal{D}t'\,\mathcal{N}\,e^{-U(Rt+\sqrt{1-R^2}\,t')}}\,, \tag{5}$$

where $\mathcal{D}t = dte^{-t^2/2}/\sqrt{2\pi}$. Since $U(b)$ is assumed to be known, $R(\alpha)$ can be determined from (5) (supplemented with the appropriate initial condition), and can be used during a practical implementation of Bayes on-line learning; cf. (4). We finally note that (5) recursively guarantees the existence of all derivatives of $R(\alpha)$, and therefore so-called phase transitions [12–14] will never occur in this learning scenario.

Before turning to a general discussion of (4) and (5), we mention the results for a few particular cases of interest. Consider supervised learning from a uniform distribution of patterns classified by a teacher perceptron. The distribution of aligned patterns is given by $\tilde{\mathcal{P}}(b) = 2e^{-b^2/2}\Theta(b)/\sqrt{2\pi}$. The results obtained from Bayes supervised on-line learning [9] are then recovered from our on-line unsupervised equations (4) and (5). Second, we consider a nonuniform distribution for which the calculations are particularly simple, $U(b) = -ab^2/2$. One finds that $F_{\rm opt} = (1 - R^2)at/(aR^2 - a - 1)$ is linear in t and $dR/d\alpha = a^2R(1 - R^2)^2/2(1 + a)(a + a - aR^2)$. This example illustrates two features that will be obtained in a more general setting below. First, on-line learning from scratch is impossible for the distribution under consideration: $R(\alpha = 0) = 0 \to R(\alpha) \equiv 0$, $\forall\,\alpha$. Second, for $R(\alpha = 0) > 0$, one finds a rather slow asymptotic approach in the $\alpha \to \infty$ limit, $R \overset{\alpha\to\infty}{\to} 1 - (1 + a)/2a^2\alpha$.

We now derive general results valid in the small or large α limit. One readily sees that the integral on the r.h.s. of (5) can be expanded about $R = 0$ as $\langle b\rangle^2 + O(R^2)$, where $\langle b\rangle = \int b\mathcal{P}(b)\,db$. For $\langle b\rangle \neq 0$ it follows that $R(\alpha) = \sqrt{\alpha\langle b\rangle^2}$ for asymptotically small α. Similarly, one finds that $F_{\rm opt} = \langle b\rangle/R$ and thus $F_{\rm opt} = {\rm sgn}\langle b\rangle\sqrt{N/p}$, showing that Bayes on-line learning (3) coincides with the Hebb rule, $\boldsymbol{J} \sim \sum_{\mu=1}^{p}\boldsymbol{\xi}^{\mu}$, in this limit. On the other hand, for $\langle b\rangle = 0$ we can conclude that $R(\alpha) \equiv 0$; i.e., any kind of on-line learning from scratch necessarily fails.

Turning to large α, one obtains, by a straightforward expansion of (5) around $R = 1$, the following asymptotic behavior:

$$R(\alpha) = 1 - \left[2\alpha\int U'(b)^2\mathcal{P}(b)\,db\right]^{-1}, \tag{6}$$

provided the integral in the r.h.s. of (6) converges. As expected, a more pronounced structure in the pattern distribution [large $U'(b)$] facilitates Bayes on-line learning. Regarding singular derivatives $U'(b)$ we restrict ourselves to a particularly interesting special case: Assume that $\mathcal{P}(b) = 0$ for $b < b_1$, $\mathcal{P}(b)$ smooth for $b > b_1$, and $\lim_{b\to+b_1}\mathcal{P}(b) = \Delta\mathcal{P} > 0$, a situation which typically occurs in supervised problems; cf. (2). By closer inspection one then finds that in the region of t and t' values which, for $R \to 1$, contribute notably to the integral in (5), $\exp[-U(\tau)]$, with $\tau = Rt + \sqrt{1 - R^2}\,t'$, can be replaced by $\exp[-U(b_1)]\Theta(\tau - b_1)$. This yields the leading order asymptotics

$$R(\alpha) = 1 - 4\pi\left[\alpha\Delta\mathcal{P}\int \mathcal{D}t\,e^{-t^2/2}H(t)^{-1}\right]^{-2}, \tag{7}$$

where $H(t) = \int_t^{\infty} Dt'$. More general discontinuities of $\mathcal{P}(b)$ lead again to an α^{-2} decay with similar coefficients. By comparison of (6) and (7), one concludes that it is easier to estimate the orientation $\boldsymbol{B}$ for discontinuous than for smooth pattern distributions at large α. Furthermore, the large-α on-line Bayes learning in the discontinuous case (7) is completely dominated by the examples close to the discontinuities of the relevant pattern distribution [i.e., $\mathcal{P}(b)$ for unsupervised and $\tilde{\mathcal{P}}(b)$ from (2) for supervised learning]. A smoothening of these discontinuities, e.g., due to a "noisy teacher," gives rise to a finite $1/\alpha$ coefficient in (6) and thus significantly undermines the progress of a "master student."

Next we address off-line Bayes learning. The calculations in this case are more involved and are only sketched, while relying on the general formalism presented in [13]. Given a training set $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1}^{p}$ it follows from the "Bayes rule" [13,14] that, for a uniform prior distribution on $\boldsymbol{B}$, the *a posteriori* probability for a hypothesis $\boldsymbol{J}$ to coincide with the unknown "true" $\boldsymbol{B}$ is proportional to $\prod_{\mu=1}^{p}P_{\boldsymbol{J}}(\boldsymbol{\xi}^{\mu})$. Sampling at random a $\boldsymbol{J}$ vector according to this distribution is known as Gibbs learning and the corresponding overlap $R_G(\alpha)$ can be obtained by a standard replica calculation [13]. According to a general argument given in [14] the maximal overlap R that can be achieved on the basis of the information contained in the training set is given by $R = \sqrt{R_G}$. This overlap is realized by the weighted center of mass of the *a posteriori* probability. After manipulations, one finally obtains the

following closed equation for this so-called Bayes off-line overlap [15]:

$$R^2 = \alpha \int \mathcal{D}t \, \frac{\{\int \mathcal{D}t' t' \mathcal{N} e^{-U(Rt+\sqrt{1-R^2}\,t')}\}^2}{\int \mathcal{D}t' \mathcal{N} e^{-U(Rt+\sqrt{1-R^2}\,t')}}, \quad (8)$$

showing a remarkable similarity with (5). While Bayes on-line learning (1)–(5) leads to the best possible hypothesis $\boldsymbol{J}$ under the extra condition of sequential updating, one may wonder under which conditions it saturates the Bayes off-line limit. One verifies that (5) and (8) are satisfied simultaneously $\forall\, \alpha$ only for the trivial case of a linear $U(b)$, resulting in Hebbian learning. Yet on-line and off-line, in fact, give remarkably similar results for a general $U(b)$ in the small and large α regimes, as we now proceed to show.

To extract the small and large α asymptotics from (8) one can use the same arguments as for (5). For small α one finds that the leading order behavior of Bayes limit on-line and off-line are identical. It follows that the Hebb rule saturates the Bayes off-line limit for asymptotically small α. This observation has been made previously for several special cases but its general validity is demonstrated here for the first time. As before, so-called retarded learning [12,16,17], i.e., $R(\alpha) = 0$ for α small, occurs if and only if $\langle b \rangle = 0$. In contrast to the on-line case, however, the off-line retardation extends only up to a finite α value [18]. For example, the previously discussed model $U(b) = -ab^2/2$ gives rise to $R(\alpha) = 0$ for $\alpha < (1 + a)^2/a^2$ and $R(\alpha) = \sqrt{[\alpha - (1 + a)^2/a^2]/[\alpha - (1 + a)/a]}$ for larger values of α; cf. also Eq. (53) in [13].

Let us now turn to the large α behavior. One finds the surprising result that for smooth pattern distributions the asymptotic behavior of the off-line Bayes limit is identical to the one for on-line; cf. (6). For discontinuous pattern distributions, on the other hand, Bayes limit off-line uses the examples "twice as efficiently," $R_{\text{off-line}}(\alpha) = R_{\text{on-line}}(2\alpha)$, with the on-line result given by (7). The latter simple relationship was already observed for several specific scenarios in supervised learning [9,11, 19–21], but remains valid in the context of unsupervised learning only for learning from a discontinuous distribution. Apparently, the asymptotic difference between on-line and off-line Bayes limits becomes more important as the learning task becomes easier. As a further illustration of this point, we mention that for a pattern distribution with a delta peak, e.g., $\mathcal{P}(b) = \delta(b - b_0)$, the off-line Bayes limit reaches $R = 1$ at the finite value $\alpha = 1$, while the approach of R to 1 is only exponential in the limit $\alpha \to \infty$ for on-line.

The α^{-2} decay in (7) for the overlap is familiar from supervised perceptron learning. In the standard student teacher perceptron supervised scenario [$f(x) = x$ and $U(b) = 0$], the aligned pattern distribution is discontinuous, and the resulting generalization error ϵ approaches 0 for $\alpha \to \infty$ as $\sqrt{1 - R} \sim \alpha^{-1}$. The slow α^{-1} approach of R to 1 for smooth distributions [cf. (6)] may look surprising in view of the Vapnik-Chervonenkis (VC) bound $\epsilon < \ln \alpha/\alpha$ [22]. This asymptotic behavior will apply in the case of a student perceptron $\boldsymbol{J}$ learning from a teacher perceptron $\boldsymbol{B}$ by examples generated with a probability distribution $\mathcal{P}(b)$ that goes sufficiently fast to zero for $b \to 0$, such that the distribution $\tilde{\mathcal{P}}(b)$ for the equivalent unsupervised problem (2) becomes smooth. There is, however, no contradiction with the VC bound because the generalization error has to be calculated with respect to the very same distribution that generates the training examples. Hence the student is evaluated mostly on easy questions (examples far from the decision plane). The decrease of ϵ becomes faster than $\sqrt{1 - R}$ and is found to obey the VC bound.

Finally, we turn to the identification of the Bayes vector as the minimum of an *ad hoc* cost function $E(\boldsymbol{J}) = \sum_{\mu=1}^{p} V(\boldsymbol{J} \cdot \boldsymbol{\xi}^\mu/\sqrt{N})$, thereby generalizing a result obtained recently for supervised learning from a teacher perceptron [20]. The overlap R between the minimum $\boldsymbol{J}$ of $E(\boldsymbol{J})$ and $\boldsymbol{B}$ can be obtained for a general potential V and a general pattern distribution $\mathcal{P}$ as the solution of the following equations for R and x [cf. Eq. (17) in [13]:

$$\int \mathcal{D}t[\lambda^0(t,x) - t]^2 X(R,t) = (1 - R^2)/\alpha\,,$$
$$\int \mathcal{D}t[\lambda^0(t,x) - t] Y(R,t) = R/\alpha\,, \quad (9)$$

with $X(R,t) = \int \mathcal{D}t' \mathcal{N} e^{-U(Rt+\sqrt{1-R^2}\,t')}$, $Y(R,t) = \int \mathcal{D}t' t' \mathcal{N} e^{-U(Rt+\sqrt{1-R^2}\,t')}/\sqrt{1 - R^2}$, and λ^0 the function of t and x that extremizes $V(\lambda) + (\lambda - t)^2/2x$. The meaning of λ^0 can be clarified by the application of the cavity method (cf. [7,23]) as the overlap of a pattern $\boldsymbol{\xi}^\mu$ with the $\boldsymbol{J}$ vector after learning this pattern as a function of the overlap t prior to learning. In view of the similar meaning of the amplitude factor F in on-line learning, it is tempting to introduce the analogous off-line quantity $F = \lambda^0 - t = -xV'(\lambda^0)$. From (9), one gets

$$\frac{R^2}{1 - R^2} = \alpha \frac{\langle YF \rangle^2}{\langle XF^2 \rangle} = \alpha \frac{\langle (Y^2/X) XF/Y \rangle^2}{\langle (Y^2/X)\,(XF/Y)^2 \rangle}, \quad (10)$$

where the average is with respect to the Gaussian measure $\mathcal{D}t$. It is easy to verify, e.g., on the basis of the Schwartz inequality, that $\langle uv \rangle^2 \le \langle u \rangle \langle uv^2 \rangle$ for any function $u > 0$, with the equality sign attained only for v constant with respect to the variables over which the average is being taken. It follows that the r.h.s. of (10), and hence also the value of R, is maximal for the choice $F_{\text{opt}} = CY/X$, where C is a constant independent of t. Its value can be determined from (9): $C = (1 - R^2)/R$. By filling in the explicit forms of X, Y, and C, one verifies that F_{opt} is identical to the result (4) found in on-line learning. Furthermore, the r.h.s. of (10) simplifies

to $\alpha\langle Y^2/X\rangle$, so that (10) becomes identical to (8), the equation for the off-line Bayes limit. Hence, we have identified a cost function E through the explicit form of $F_{\rm opt}$, whose unique minimum reproduces the off-line Bayes result. The identity of $F_{\rm opt}$ for the Bayes limit on-line and off-line sheds light on the relation between both solutions. The Bayes limit off-line is such that, upon removing any of the learned patterns, the relearning of that pattern is optimal in the on-line sense. Furthermore, this is true for all of the learned patterns in off-line, whereas it only holds for the last pattern in on-line. We finally describe how the explicit form of the optimal cost function can be constructed. The function $\lambda^0_{\rm opt} = F_{\rm opt} + t$ is monotonically increasing function of t, going from $\lambda_{\rm min}$ to $\lambda_{\rm max}$, for any value $R \in]0, 1[$ and any choice of $U(b)$. Hence, the inverse function $t_{\rm opt}(\lambda)$ is well defined in the interval $I = [\lambda_{\rm min}, \lambda_{\rm max}]$. We construct a potential $V_{\rm opt}$ by integration of $V'_{\rm opt}(\lambda) = [t_{\rm opt}(\lambda) - \lambda]/x$ for $\lambda \in I$ and $V_{\rm opt}(\lambda) = +\infty$ otherwise. The choice of the proportionality factor x is related to the speed of covergence of a gradient descent algorithm on the cost function E. Since it does not modify the location of the minimum of E, its value is left undetermined. It is clear from the above construction that, with a few exceptions like a linear $U(b)$, $V_{\rm opt}(\lambda)$ depends on α, and closed analytical expressions can only be obtained in special cases. However, the small and large α behavior can be extracted, revealing which learning strategies are Bayes optimal in these limits. For asymptotically small α one finds $V_{\rm opt}(\lambda) = -\lambda$, which is equivalent to the Hebb rule, is agreement with out previous results. For asymptotically large α and smooth pattern distributions $\mathcal{P}(b)$ governed by (6), $V_{\rm opt}(\lambda) = U(\lambda)$, so that minimizing E becomes equivalent to maximizing the *a posteriori* probability $\prod_{\mu=1}^{p} P_{\boldsymbol{J}}(\boldsymbol{\xi}^\mu)$ [13,14]. In other words, maximal likelihood learning is Bayes optimal in the large α regime.

Discussions with N. Caticha are gratefully acknowledged. We thank the Program on Inter-University Attraction Poles of the Belgian Government, the NFWO Belgium, and the Holderbank Foundation (Switzerland) for financial support.

*Present address: Eötvös University, Puskin-u. 5–7, H-1088 Budapest, Hungary.

[1] G. Györgyi and N. Tishby, *Neural Networks and Spin Glasses,* edited by W. Theumann and R. Koberle (World Scientific, Singapore, 1990).

[2] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computing* (Addison-Wesley, Reading, MA, 1991).

[3] H. S. Seung, H. Sompolinksy, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).

[4] T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).

[5] M. Opper and W. Kinzel, in *Physics of Neural Networks III,* edited by E. Domany, J. L. Van Hemmen, and K. Schulten (Springer, Berlin, 1994).

[6] C. Van den Broeck, Act. Phys. Pol. B **25**, 903 (1994).

[7] M. Bouten, J. Schietse, and C. Van den Broeck, Phys. Rev. E **52**, 1958 (1995).

[8] A. Engel, Mod. Phys. Lett B **8**, 1683 (1994).

[9] O. Kinouchi and N. Caticha, J. Phys. A **25**, 6243 (1992).

[10] M. Copelli and N. Caticha, J. Phys. A **28**, 1615 (1995).

[11] M. Copelli, O. Kinouchi, and N. Caticha, "Equivalence Between On-Line Learning in Perceptrons with Noisy Examples and Committee Machines," Universidad de Sao Paulo, 1995 (to be published).

[12] M. Biehl and A. Mietzner, Europhys. Lett. **24**, 421 (1993); J. Phys. A **27**, 1885 (1994).

[13] P. Reimann and C. Van den Broeck "Learning by Examples from a Non-Uniform Distribution" [Phys. Rev. E (to be published)].

[14] T. L. H. Watkin and J.-P. Nadal, J. Phys. A **27**, 1899 (1994).

[15] The derivation of (8) relies on the assumption that replica symmetry (RS) holds in the Gibbs scenario. As a justification we mention that the local stability condition of this RS solution was found to hold true for all special cases considered so far and can be verified explicitly for general pattern distributions (1) provided α is asymptotically small or large.

[16] E. Barkai and I. Kanter, Europhys. Lett. **14**, 107 (1991).

[17] G. J. Bex, R. Serneels, and C. Van den Broeck, Phys. Rev. E **51**, 6309 (1995).

[18] A detailed discussion of this point, including a stability analysis of the replica symmetric solution and the possibility of first order phase transitions, will be given elsewhere.

[19] M. Biehl, P. Riegler, and M. Stechert, "Learning from Noisy Data: An exactly Solvable Model," University of Wuerzburg (to be published).

[20] O. Kinouchi and N. Caticha, "A Learning Algorithm which Gives the Bayes Generalization Limit for Perceptrons," Universidad de Sao Paulo, 1995 (to be published).

[21] It has also been proven for supervised learning in a $K = 2$ parity machine with nonoverlapping fields [N. Caticha (private communication)].

[22] J. M. R. Parrondo and C. Van den Broeck, J. Phys. A **26**, 2211 (1993).

[23] M. Griniasty, Phys. Rev. E **47**, 4496 (1993).